Relativity

EBOOK

How Reliable is Your AI? The Legal Team's Guide to AI Evaluation

A practical framework for choosing, testing, and continuously evaluating legal AI



Contents

Executive Summary 3

Section 1: **The Checkbox Trap 4**

- Adoption without Measurement is Doomed 5
- The Unique Opportunity for Successful AI Adoption in Legal 6

Section 2: **Key Concepts to Know and Share 7**

- Key Terms 7
- The Human in the Loop 9
- Broad Benchmarking versus Use Case Evaluation 9
- Validation as a Form of Evaluation 11
- Capability versus Reliability versus Accuracy 11
- Understanding Fit for Purpose 12

Section 3: **The Accountability Question 13**

- Accountability Always Falls on Human Shoulders 14
- Correcting for Personal Biases toward AI 14
- Discovery and Disclosure: Where Proactive Expectations Emerge 15

Section 4: **An AI Evaluation Framework 16**

- The 5C Evaluation Framework for Legal AI 17
- Stage One: Claim 18
- Stage Two: Context 19
- Stage Three: Calibrate 20
- Stage Four: Conduct 22
- Stage Five: Continue 24

Section 5: **Evaluation Insights from the Field 25**

- How Practitioners Choose Tools 26
- Questions to Ask Any AI Vendor Before Adoption 27

Section 6: **What the Industry Hasn't Settled 29**

- AI Evaluating AI 30
- Tolerance of Output Variability 30
- The Necessity of Shared Industry Evaluation Standards 31

Section 7: **Planning for the Future 33**

- Evaluation as a Defensibility Record 34
- Evaluation Maturity Reflects AI Maturity 34

Appendix: **Related Resources to Keep Exploring ... 35**

Executive Summary

Legal teams are grappling with a dangerous problem: AI's capabilities often outpace its reliability and economy at scale. What a system can do in a controlled demonstration and what it does consistently, on your data, for your team, and under your circumstances, don't always overlap very neatly – and many legal AI users aren't yet measuring that gap.

Meanwhile, a lawyer's professional duty of competence remains as binding as ever, even when technology is new and innovating rapidly. The consequences for deploying a tool a user may not fully understand, accepting its outputs without scrutinizing them, or assuming the product's own benchmarks are good enough to validate your use case can be swift and they can be devastating.

At the heart of this crossroads is a hard truth: evaluating AI – thoughtfully, specifically, consistently – is every legal practitioner's responsibility.

Owning this truth will help you avoid those consequences and make the most of your technology investments. But fear not: the skills required are closer to hand than you think.

Collectively, the legal community fostered a culture of skepticism, validation, and quality control years before evaluations entered the broader enterprise AI vocabulary. The discipline that made technology-assisted review defensible is the same discipline which can help separate AI deployments that deliver from those that introduce untenable risk or simply exist to check a box.

The work is learning how to apply that muscle – continuously! – to a new generation of tools.

This e-book is designed to help.

Inside, you'll hear from litigators, legal operations leaders, data scientists, and service providers who are doing this work now, on real tools and projects. We also offer the 5C AI Evaluation Framework – Claim, Context, Calibrate, Conduct, Continue – as a practical structure for teams ready to operationalize their legal AI evaluation strategies.

The teams that thrive in this era will be the ones who both embrace AI and know how to show their work. Ready to join them?

Disclaimer: The views expressed in the quotes featured throughout this piece are those of the individual speakers and do not necessarily represent the views of their respective organizations or any of their clients.

Section 1: The Checkbox Trap

The pressure to adopt AI across enterprise categories is immense.

From budget-minded leaders to innovation-hungry clients and eager-to-learn professionals occupying the many roles in between, the narrative around AI's potential to transform modern work has taken hold.

> "I'm not experiencing much pushback on using AI anymore. People understand that there has been some battle testing of these tools," Liz Gary, of counsel at Morgan Lewis and a 2026 AI Visionary, observes. "This is a technology that, when you have a trusted advisor recommending it, you can accept it. And I think that's because, people are using this technology more and more, at work or in their personal lives."

©Relativity. All rights reserved.

As a result, maturing expectations and requests for AI usage are pouring in across segments:

- Where are you using AI to improve your efficiency?
- How can AI help us level up this workflow or deliverable?
- Which AI tools have you onboarded so far?
- Are you training your team on AI fluency?

For some folks, this is energizing stuff. Among these enthusiasts, investments in brand-new AI tools promising innovative takes on stubborn challenges have come rolling in.

For others, it's ... motivational, perhaps, but still rouses hesitation. Many of these teams get started once artificial intelligence features become embedded into enterprise tools they already know and trust.

Either way, AI adoption has gone mainstream. In a **McKinsey survey** from late 2025, "almost all respondents say their organizations are using AI, and many have begun to use AI agents."

Impact, however, is a different story. **The Boston Consulting Group's 2026 AI at Scale study**, which surveyed 1,800 executives, found that only 26 percent of companies reported seeing meaningful financial value come from their AI investments.

So what gives? How can enthusiasm and investment be so high, while ROI is so low?

We like to call it the "checkbox trap."

Adoption without Measurement is Doomed

For teams looking to get real value out of their AI investments, adoption without evaluation is meaningless. Most organizations are onboarding AI, but not all of them are systematically checking whether it performs reliably at the scale required – for whom, on which tasks, under what conditions, and whether it stays useful as tools, data, and needs evolve.

The checkbox gets ticked, and for a bit, people are happy. We have AI! It's nifty! But the rigor doesn't follow, and that's where those investments fall flat – or worse.

"We see so much out in the wild that indicates that people are not evaluating, people are not validating," says Aurora Bryant, a senior manager of applied legal intelligence at Relativity. "There are over a thousand court decisions now worldwide – it might even be in the many thousands by now – of lawyers citing cases that don't exist, or that do not support the proposition for which they are cited."

This is the risk legal teams encounter when using AI without scrutinizing it.

"The first time that this happened, everybody thought, 'oh, this is never gonna happen again,'" Aurora continues. "But it did, and it does, which tells me that some teams are not evaluating their system for their use case, they're not validating their results, or they're not doing either."

So poor evaluation – up front and over time – threatens poor legal outcomes, reputational damage, and wasted resources. The ripple effect doesn't end there. A breach of trust in the technology may take ages to repair, which means a poorly executed AI pilot could muddy the waters for other, better adoption plans for a long time. When the tech is evolving this rapidly, it's painfully easy to fall behind the curve (and your competition) while you wait for those fears to soften.

Meanwhile, developers are making real technical progress on problems like hallucinations – but no one claims hallucinations (or other potential pitfalls) have been eliminated. Enthusiasm is justified, but skepticism and caution remain warranted. That's where proper evaluation comes in.

©Relativity. All rights reserved.



The Unique Opportunity for Successful AI Adoption in Legal

This all sounds bleak, but there's good news, too: legal is uniquely positioned to buck these trends and facilitate successful, productive, and safe AI pilots.

This is because the defensibility culture inherent to this field gives us a leg up on evaluation as a discipline – one that was established long before generative and agentic AI came along to challenge it.

Legal data intelligence use cases, especially e-discovery, demanded validation, sampling, and statistical quality-control norms a decade before "evals" became an AI buzzword. In the bygone era of the early 2010s, technology-assisted review workflows came in to help address the explosive data volumes that weighed heavily on legal teams conducting litigation, investigations, records requests, and other legal data projects.

Plenty of rigor was put into testing, validating, and defending these early AI tools. Today, most legal professionals tasked with large-scale data review projects already understand concepts like precision, recall, and overturn rates.

"This moment in AI is a rare opportunity for legal to show we are not just adopting technology: we're activating innovation within legal itself," says Christine Hasiotis Martins, a 2026 AI Visionary and global e-discovery leader and senior litigation counsel for GE Vernova. "This is our chance to connect curiosity with experimentation in ways our function has never experienced before. It is our chance to ask 'what if?' and 'why not' rather than defaulting to 'we've always done it this way.'"

Section 2: Key Concepts to Know and Share

Before you move into more tactical evaluation conversations around modern AI, you've got to make sure all stakeholders are speaking the same language.

This sounds like a given but, in practice, a surprising number of AI conversations get off track simply because of inconsistent lingo. A litigator asking whether "the AI" is "reliable" and a vendor explaining that "the model performs well" may both be speaking in good faith but also be talking about entirely different things.

A shared vocabulary can help you shape questions, understand and challenge their answers, and take ownership of the accountability that still falls on ***humans*** if something goes wrong.

Key Terms

This section establishes a couple of key words and concepts that the rest of this e-book depends on. To begin, Aron Ahmadia, vice president of applied science at Relativity, describes some definitions for us.

©Relativity. All rights reserved.

Foundation Model

"A foundation model is a pre-trained large language model; I think we've all heard this term more times than we wanted to in the last three to four years. But when we talk about a large language model, or sometimes when we just say 'model,' we're referring to that **foundation model** underneath a tool," Aron explains. "There are a bunch of layers between you and it, so you don't usually interact with it directly; you're talking to a system."

In the context of evaluation, this layering means "model capability" and "system performance" are related but distinct concepts.

Harness

From that foundation, "we'll climb that stack of layers to the harness, which is the set of composing tools that either a provider or vendor has built around that large language model to help it accomplish tasks on your behalf."

Two teams running the same model in different **harnesses** may see very different results. That's by design, but it's a good snapshot of why "AI evaluation" is a more subjective concept than you might guess.

AI System

"Next," Aron says, "take one more step up to the **AI system**. An AI system is, in some sense, a complete solution containing a large language model that is equipped with a harness."

In other words, this is the full stack: the model plus the tools, workflows, and integrations layered around it. Evaluating a "system" is fundamentally different than evaluating a model in isolation.

Evaluation

"An **evaluation** is something that measures a claim, and it can be many, many things. It could be rigorous, it could be informal, it could be ongoing, but the point is that somebody has made a claim and an evaluation is a way to measure against that claim," Aron says.

"So the first step," Aron advises, "is knowing what claim you are making."

In the context of legal work, that answer might lie in some familiar questions: Can I trust this output? Does this system make me more productive? Does it meet the standard for a defensible process?

©Relativity. All rights reserved.

The Human in the Loop

"Human judgment absolutely matters," says Mahendra Marshall, associate director of project management at Cimplifi. "That's generally not optional for legal use cases."

For legal teams, the input of human experts throughout your AI use and evaluation is non-negotiable. There are ways to have AI evaluate itself, and they can be helpful at scale, but they can't be used in lieu of thoughtful scrutiny of the data they produce and how they arrived at their conclusions.

You've heard that keeping a **human in the loop** on the execution of AI is essential; it's the same for the evaluation of AI.

For example, in some legal workflows – such as contract or brief drafting – every AI output should be checked by a human. Others are performed at such a scale that absolute perfection may not be the standard – for example, document review – and so checking every single output becomes unfeasible.

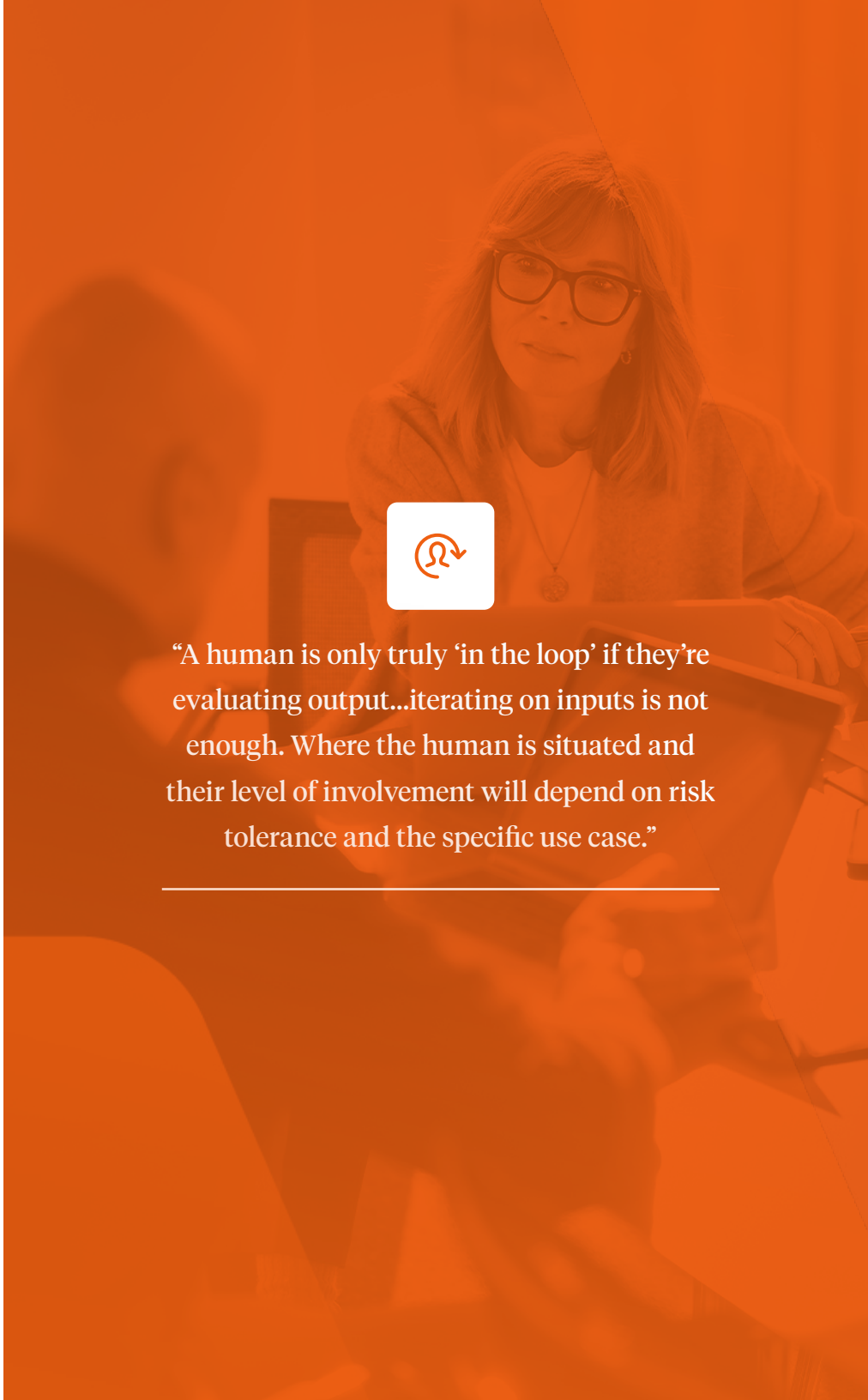
Broad Benchmarking versus Use Case Evaluation

Sometimes you'll want to evaluate AI broadly:

- Is this tool well suited for my team and the type of work we do?
- Is the model responsibly trained?
- Is the system secure?

But other times, you'll need to get more specific:

- Are its answers accurate?
- Does it perform consistently on this type of project?
- How often do humans have to correct its work?
- What does it do when it doesn't know the answer to a user's question?



"A human is only truly 'in the loop' if they're evaluating output...iterating on inputs is not enough. Where the human is situated and their level of involvement will depend on risk tolerance and the specific use case."

©Relativity. All rights reserved.

For broad questions, you might reference **benchmark** data provided by the developer and/or third-party evaluators. Vendors who freely and transparently provide their own testing details, references from other customers, and/or third-party validation are waving green flags.

"You don't have to be technical to hold AI providers accountable. The key is asking simple, practical questions about how the technology works, how results are validated, and what safeguards are in place if something goes wrong," **advises AI Visionary Andrew Milauskas**, who is chief operating officer at Page One. "Trust comes from providers being open about what their AI can and cannot do as well as their willingness to stand behind the results."

(Note: We share more on what questions to ask potential AI vendors in Section 5.)

Benchmarks are helpful, but they may not be one-to-one with your data and tasks. That's where more **targeted, use-case evaluation** comes in.

"There are many different use cases for AI, and developer evaluation is, of necessity, incomplete in many situations in the real world. We do our very best to test on a good simulation of the real world, but we may or may not know, for example, if a slightly different specific use case will come about," explains Aron Ahmadia.

In such cases, the testing has to come down to you.

"You need to consider doing both ongoing broad evaluation and specific use case evaluation," notes Mahendra. "At Cimplifi, we have a support team that is analyzing systems before they come to market to make sure they're safe and secure as a form of broad evaluation. But we also have review consulting teams engaging with clients to define their end goals, select systems, and conduct validation for their specific use cases so that we have focused evaluation ongoing as well."

Rishi Chhatwal, a partner at Redgrave, agrees.

"You need to have some broad evaluation up front so you know that the tool or model that you're using is fit for that purpose. But it has to continue to happen – as purposes change, or if there are updates to the model or the system it resides in," he says. "This is another situation in which you might see that even a great tool used all over your organization may not be fit for the purpose that you need it for. In the legal context, that might mean your outside counsel can't rely on it without conducting their own evaluation to ensure they can validate it in their context, too."



Benchmarks are helpful, but they may not be one-to-one with your data and tasks. That's where more targeted, use-case evaluation comes in – and that testing comes down to you.

©Relativity. All rights reserved.



Validation as a Form of Evaluation

Another quick keyword to keep in mind: **validation.**

"Evaluation is something that you can do at any point. It can be informal, it can be formal, rigorous, cheap, expensive, slow, fast. To do an evaluation is just to measure in some way," Aron explains. "But a validation is a specific type of evaluation, typically for a given use case. For example, if I am turning over a select set of documents to the court or an opposing party, I'll explain how I have validated my search for that select set: I reviewed the results, or I took a statistical sample, or something along those lines, to confirm that the documents that I said were not relevant for this matter actually weren't, and I didn't miss too many of the ones that were."

Again, the type of validation you conduct will depend on the needs of each use case.

"Differing risk tolerance means that the validation on a particular matter and the use of a particular tool may be different," Rishi agrees. "You might need to validate to a much lower level of scrutiny if it's an internal investigation or a higher level if it's something you're going to rely on in court."

"You evaluate the system, you validate the results," says Aurora Bryant. "These ideas of evaluation and validation are different, but they're both necessary."

Capability versus Reliability versus Accuracy

Sometimes the terms capability, reliability, and accuracy are used interchangeably in everyday conversation about evaluating AI. We suggest they shouldn't be.

Accuracy is correctness – freedom from mistakes or errors. If AI can give you an accurate answer to a question, in an appropriate format, that shows its **capability** of meeting your requirements.

"Capability is the question of whether the tool can do the thing," says Aurora Bryant. "A Q&A model that produces a correct, well-cited answer one time has demonstrated capability."

Reliability, on the other hand, is defined by whether that AI can be expected to capably deliver accurate outputs at the volume and consistency your use case demands.

"If you ask that same question in 50 separate attempts and receive an equally confident but factually inaccurate answer 50 percent of the time, that tool has a serious reliability problem," Aurora continues.

So, if AI can give you an accurate answer to a single question, that means it is capable of doing the task you've given it. More importantly, though: if AI can demonstrate that capability consistently over time, at the volume and within the margin for error you'd consider appropriate for your use case, you'd call it "reliable."

Understanding Fit for Purpose

You probably noticed a lot about "you" in that last paragraph. This is where that subjectivity we mentioned really comes out.

If you are selecting an AI system, it isn't enough to ask if it's accurate. You'll need to be prepared to evaluate its capability and reliability within the use case you're hoping to tackle – ensuring it's fit for purpose within the context of your work. That means it's specifically engineered to support the types of work you plan to do with it.

"AI is not a magic solution to every problem, and not all generative AI tools have the same set of strengths and weaknesses. From experience, I can tell you that you can drive a nail with a crescent wrench, but a hammer is better. Depending on the type of nail or the material you are nailing, a specific type of hammer might even be better. And if you need to hammer 15,000 nails, a pneumatic nailer may be the only reasonable approach," explains Deric Yoakley, lead counsel of e-discovery at Walmart.

"Take the time to examine and understand both the tool and the need. If you are worried about hallucinated case citations in AI-written briefs, don't use the tool to do so unless you are willing and able to invest the time and guardrails necessary to own and trust the output," he continues. "If that risk still gives you pause, fine. Think about dozens of other use cases in your practice where these tools may be able to help you be better and faster. I can assure you your competition is acting now."

Deric's sense of urgency is justified, as is his advice about finding the approach that feels right for you. Even after you've selected and settled into a tool, ongoing evaluation remains subjective; it's got to be rooted in the tasks, projects, risk tolerances, and data of your team and clients.

Your evaluation approach will be uniquely yours – which means it'll be golden for getting you the information you want, but also that it'll take some work to build.

©Relativity. All rights reserved.

Section 3: The Accountability Question

One of the most fascinating things about legal's quick embrace of generative AI is how it has demonstrated that, as an industry, folks seem to agree we don't need to formally measure these tools at all times to benefit from their use for many tasks.

Lawyers are responsible for their work product whether they've used AI to help create it or not – that's a common-sense assumption, but it's also emphasized by ethical obligations and, in jurisdictions like California, the law. For drafting, research, and similar use cases, we assume they're checking for accuracy. If not, they are accountable for any consequences earned by those inaccuracies.

Put another way, **responsible use of legal AI means understanding that the human using it is accountable for validating its outputs.** They will shoulder the consequences of any errors.



Lawyers are responsible for their work product whether they've used AI to help create it or not – that's a common-sense assumption, but it's also emphasized by ethical obligations and, in jurisdictions like California, the law. For drafting, research, and similar use cases, we assume they're checking for accuracy. If not, they are accountable for any consequences earned by those inaccuracies.

©Relativity. All rights reserved.

Accountability Always Falls on Human Shoulders

Make no mistake: "You can't offload accountability from the human," Mahendra Marshall emphasizes. "So there is no world in which you can remove human judgment completely from the equation."

Legal professionals' preparedness to take on this accountability is a win for the field when it comes to AI adoption. These check-your-work reflexes are already woven into their everyday habits.

"You don't have to touch every output manually," Mahendra continues, adding to the case-by-case nature of evaluation needs. "You could do sampling, or focused review where you feel risk is highest, or validation at key points – but human judgment is still required."

Correcting for Personal Biases toward AI

An important step for those humans is to address the natural biases that cloud their judgment around the reliability, defensibility, and useability of AI. The healthy skepticism that underlies all good evaluation habits grows from here.

Automation bias is one example. It's the natural tendency humans have to trust a machine's instructions or conclusions more innately than a person's, because we assume a machine is more likely to be correct.

Consider, for example, **an experiment** in which a group of people are together inside a building when a simulated fire breaks out. These subjects couldn't see the flames, but they could see smoke and hear the fire alarms. A robot appeared and offered instructions it claimed would guide them to safety. The instructions led the subjects away from marked, safe exits – but participants tended to follow them anyway. They trusted the machine, which ultimately led them nowhere.

"Lawyers who are eager to demonstrate they are using AI may over-rely on the AI output without applying the critical thinking skill that defines our legal profession. Our most valuable skills as lawyers are judgment, reasoning, questioning, verifying, and validating. It's too easy to take the output of AI, scan the format, and use it as-is because we are so impressed with how quickly the output makes us happy," observes Christine Hasiotis Martins. "We need to train our legal teams in how AI works (it is a predictor, not a reasoning tool), and in particular, develop training that demonstrates how our work changes depending on the tasks we assign to an AI tool."

Good training, as Christine notes, is key to empowering legal teams with AI in a measured, less biased, and more defensible way.

On the other hand, some legal practitioners and clients may want to dismiss the use of AI completely; their extreme caution and preference for "the way we've always done things" close off their willingness to experiment with the innovation and optimized workflows that help address modern legal data challenges.

"Some clients are naturally hesitant, particularly when the work is high-stakes," Andrew Milauskas says on this topic. "My team finds it to be very impactful when we can demonstrate the technology's accuracy against data pre-coded by humans, allowing us to quantify the results. We start small, establish workflows, and set clear expectations and outcomes. Once clients see the results and understand that AI supports human judgment, their confidence grows quickly."

Naturally, the effort you put into evaluating your AI tools and workflows can help you advocate for them with those who resist change.

©Relativity. All rights reserved.

Discovery and Disclosure: Where Proactive Expectations Emerge

As Andrew notes, many teams encounter resistance to AI adoption specifically in high-stakes, high-volume legal data intelligence use cases like e-discovery, M&A exercises, data breach responses, and e-disclosure. Here, expectations of reliability are more proactive.

Again, this healthy skepticism makes a lot of sense. “Ethics rules in the US and elsewhere establish that lawyers have a duty to their clients of technological competency (or to work with those who have it),” Rishi Chhatwal says. “This places some burden on practitioners to learn and understand the use and limits of AI.”

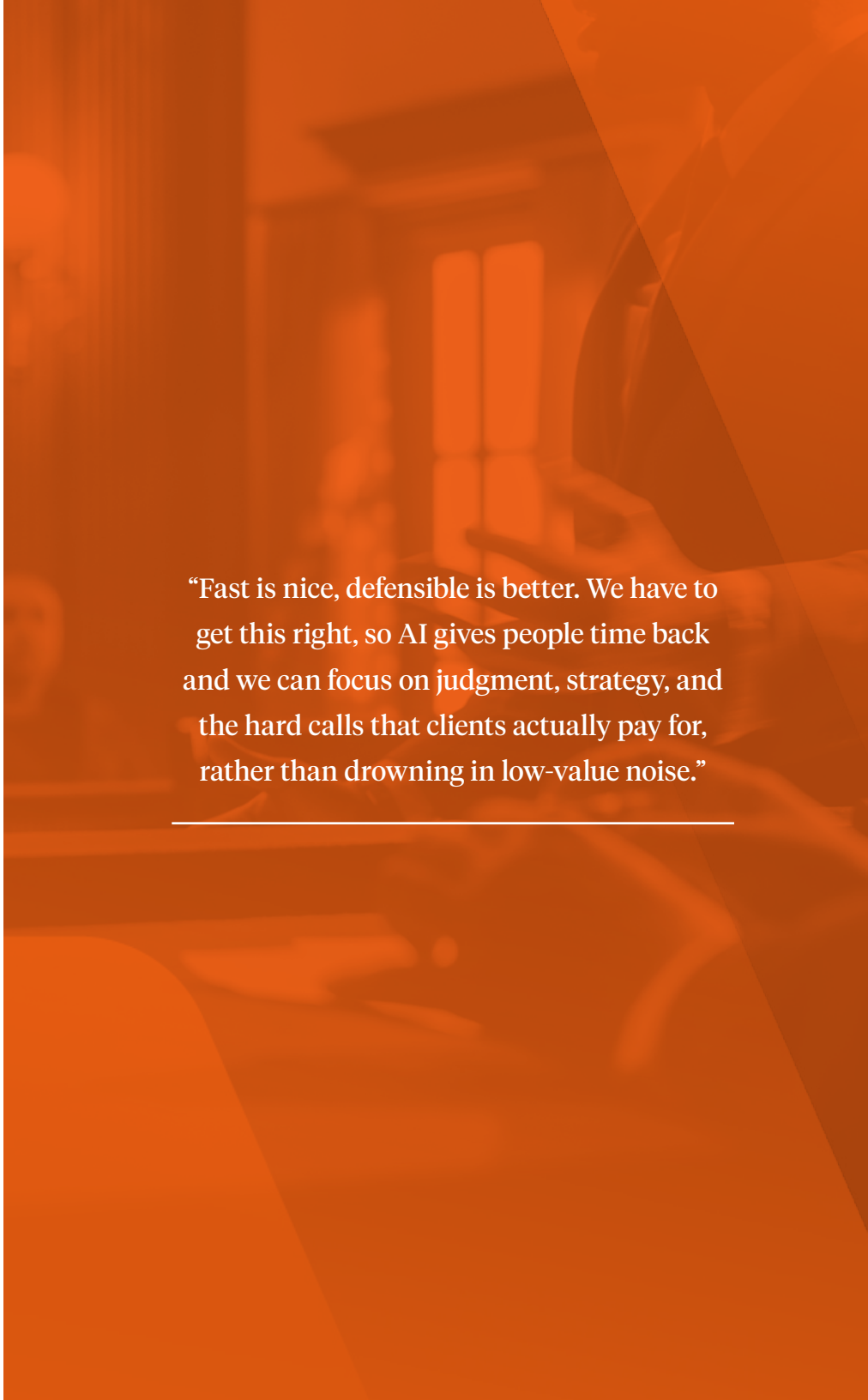
Just as those duties mean checking citations and validating claims in AI-generated briefs or motions, they require practitioners to ensure the defensibility and accuracy of a tool or workflow that’s helping a review team code thousands upon thousands of documents as related or unrelated to a litigation matter. In the latter circumstances, missing something critical can change the entire outcome of a case – so, as they’ve always had to do with more traditional review workflows, counsel must be ready to defend the accuracy of their work.

In the near-term, when AI is the everyday means teams use to accelerate these types of projects, “the differentiator will not be who uses AI; it will be who can show, clearly and defensibly, how it was used,” says Ankur Malik, global head of forensic technology and e-discovery at Clifford Chance and an AI Visionary. “Traceability will be nonnegotiable. More data, more automation, more need for disciplined process.”

He continues: “Fast is nice, defensible is better. We have to get this right, so AI gives people time back and we can focus on judgment, strategy, and the hard calls that clients actually pay for, rather than drowning in low-value noise.”

This “traceability” requirement, as Ankur puts it, explains the higher level of scrutiny required for e-discovery and legal data platforms that will be relied upon for analyzing and producing large amounts of data. This type of evaluation must be built directly into your workflows – not left to individual diligence.

©Relativity. All rights reserved.



Section 4: An AI Evaluation Framework

Alright, enough theory. You get it, right? Evaluation is important. It's also wildly different from team to team.

So then: how do you do it?

Going back to what Aron Ahmadia says: "An evaluation is something that measures a claim. And so, the first step is knowing what claim you're making."

Setting this context and articulating the goals and obligations you have in mind will help determine your evaluation workflow.

To put this into more tactical terms, we built the 5C Evaluation Framework for Legal AI: a five-stage approach designed to help legal teams scrutinize new tools, validate existing ones, and build evaluation into the rhythm of their work.

©Relativity. All rights reserved.

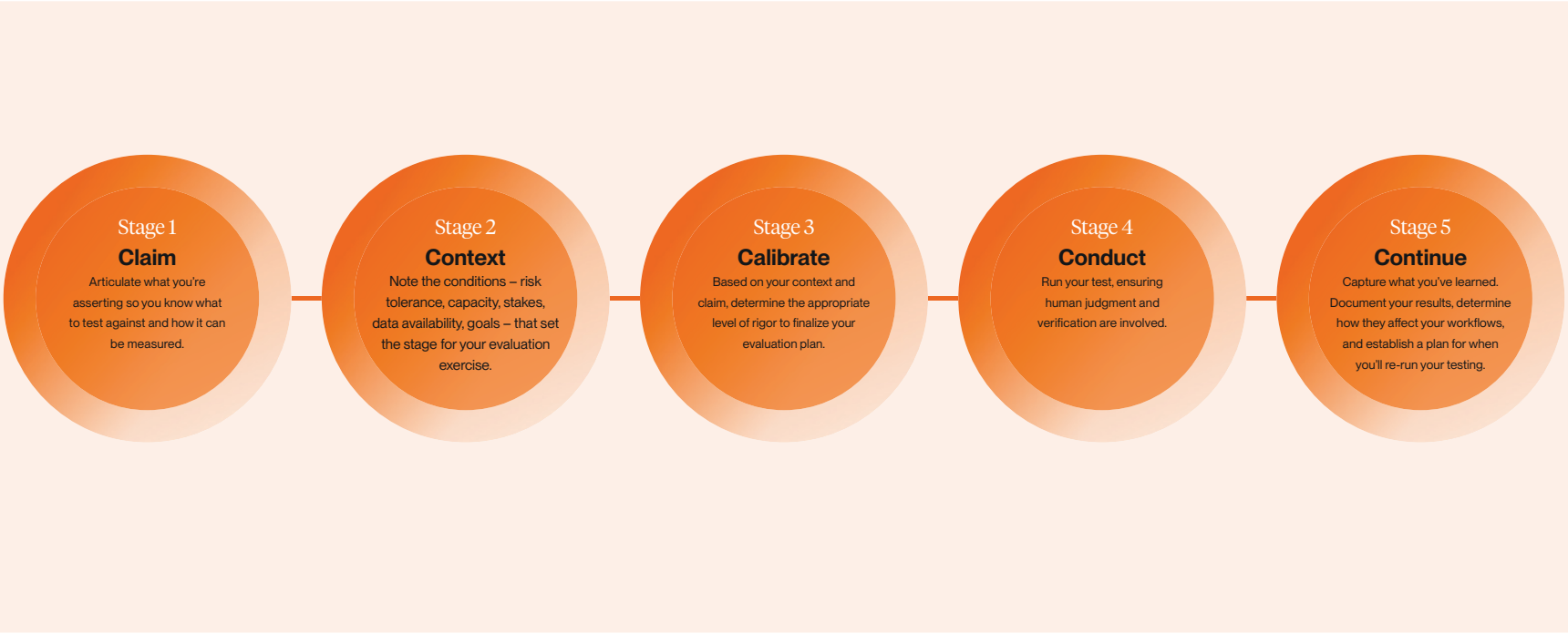
The 5C Evaluation Framework for Legal AI

Each of the five stages is intentionally straightforward, leaving room to meet the needs of each use case while providing a consistent structure that will help you strengthen these muscles over time.

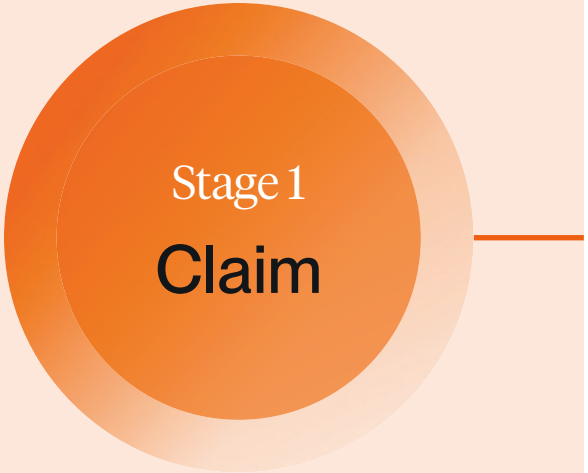
The five stages are:

1. **Claim:** Articulate what you're asserting so you know what to test against and how it can be measured.
2. **Context:** Note the conditions – risk tolerance, capacity, stakes, data availability, goals – that set the stage for your evaluation exercise.
3. **Calibrate:** Based on your context and claim, determine the appropriate level of rigor to finalize your evaluation plan.
4. **Conduct:** Run your test, ensuring human judgment and verification are involved.
5. **Continue:** Capture what you've learned. Document your results, determine how they affect your workflows, and establish a plan for when you'll re-run your testing.

Think of crafting a smart AI evaluation plan like you'd craft a really good AI prompt: the more context, intentionality, and refinement you include, the more useful your results will be.



©Relativity. All rights reserved.



Stage One: Claim

The first stage of the 5C Framework asks you to make your hypothesis explicit before you run a single test.

Starting an evaluation with "does this AI work?" is too broad; you're not likely to get a meaningful answer. Instead, craft something more specific. You can start by thinking through more focused questions:

- Does this tool accurately identify responsive documents in a data set with characteristics like ours?
- Does this assistant reduce the time our team spends on first-draft research by a measurable margin?
- Is this system's output robust enough to defend in a legal setting?

In legal, the questions you find yourself asking tend to organize around three focus areas:

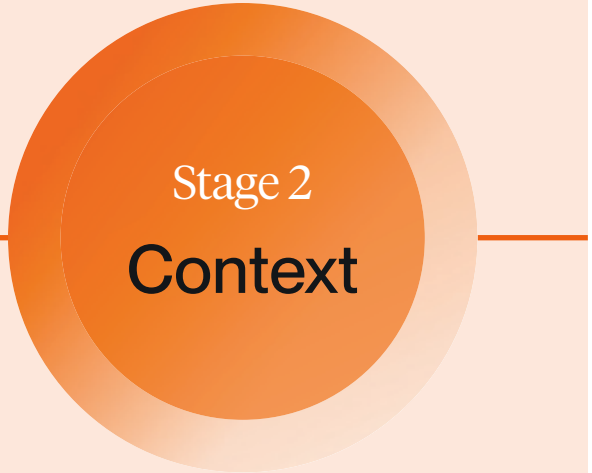
- **Trust:** Is the output accurate, well-grounded, and free of the kind of confident errors that create professional risk?
- **Productivity:** Does this system make the work faster, more thorough, or less expensive – in a way that is specifically useful to our team and can be validated?
- **Defensibility:** Does this process meet the standard a court, regulator, or client will hold it to?

Your claim is a hypothesis – it comes from your gut response to the question you need to answer. Based on your first-hand experience, what the vendor is telling you, or the quality of outputs you've seen so far, you'll craft a hypothesis you can put to the test.

Stating your claim will determine what evidence you're looking for, what methods you'll use to find it, and what result would confirm or change your supposition.

- A team testing for trust might want to verify the accuracy of AI's research or check that its claims make sense against the backdrop of their legal reasoning and the details of a case.
- A team testing for productivity will need a methodology for verifying time, cost, or risk savings from an AI-powered workflow compared to a manual one.
- A team testing for defensibility needs statistical validation of their AI's document coding, data conclusions, or case law citations.

Once you name your claim, the measurable metrics you need come into view.



Stage Two: Context

Two teams might make the same claim and run different evaluations because – you guessed it – their context differs.

In Stage 2, you note the contextual realities that will shape what a meaningful evaluation looks like for your team:

- The risk you're willing to accept if the tool underperforms
- The capacity you have to conduct research and run tests
- The data you have to work with
- The use case you're testing
- The goals you're trying to achieve by adopting a given AI system

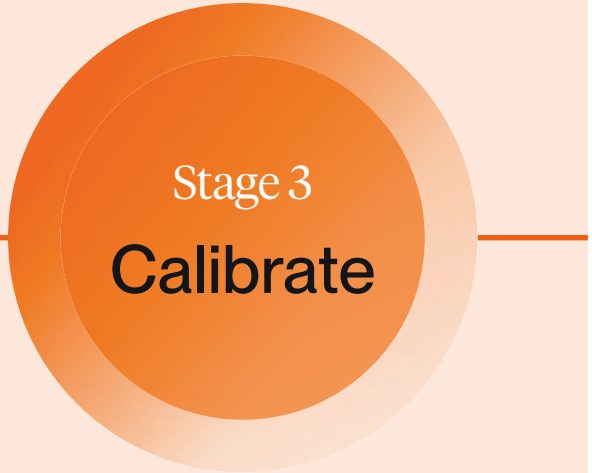
All of these color your go-forward strategy.

For example, balancing your risk appetite and your capacity might help you decide whether the scale of evaluation you need to do is something that requires help from a trusted partner.

The data and use case you're working with might have been **well-tested by someone in a similar position**, meaning their conclusions can inform your own – or, then again, your data might be niche enough to require its own validation exercises.

The Context stage is where you inventory what's true about your situation before you decide how to test your claim. It's also where you surface the constraints that will shape your methodology – including some that practitioners often underestimate, like how much time the team has to run tests, whether you have access to a representative data set, and whether research and third-party data are sufficient to meet your needs.

The discussions you'll need to have to answer these questions and articulate your context will be invaluable to each evaluation as well as your broader AI adoption strategy.



Stage Three: Calibrate

While we've established that not every evaluation needs to be a scientific study, it remains true that not every evaluation can afford to be a gut check, either.

The Calibrate stage is where you match the rigor of your evaluation plan to the risk involved – essential work if you hope to resist the temptation to either over-engineer low-stakes testing or under-invest in high-stakes validation.

This calibration is partially grounded in the accountability concepts described in Section 3 of this e-book. At the lower-stakes end of the spectrum, the individual using the tool is the main evaluation checkpoint – responsible for checking the output before acting on it, and accountable for the consequences if they don't. On higher-stakes use cases, evaluation needs to be built into the system itself: continuous, documented, and defensible.

The goals and conditions from your Context stage will help you calibrate your specific approach case by case.

Here are a few examples of what different types and scales of evaluation protocols might look like for different use cases:

If you're evaluating an e-discovery or e-disclosure tool:

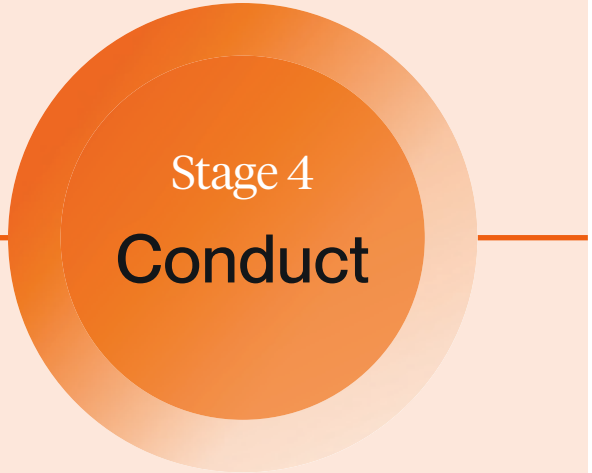
- ***For a specific project*:** Run a statistical analysis of recall and precision using a representative sample of your actual data set. If these calculations are built into your tool of choice, this checkpoint is simple – a relatively inexpensive calibration. Set a defined confidence interval to validate your results and illustrate the defensibility of your production.
- ***For general tool validation:*** Run a comparative study – AI-supported and human review on the same data set – and compare results to establish a like-for-like baseline. If you don't have the time or resources for an independent study, see if a trusted service provider can do it for you – or if other firms similar to yours have any publicly shared case studies you can review.
- ***For ongoing optimization:*** Gather statistical metrics from multiple projects into an ongoing documentation exercise, checking whether your tool serves cases well over time or your workflows need refinement. If you work with a service provider, they can do this work for you and incorporate it into touchpoints like quarterly business reviews – or, if you have the tools and time yourself, you can use an approved AI assistant to help you run your own numbers.

If you're evaluating a drafting or legal research tool:

- ***For groundedness:*** Conduct a structured citation verification review on a sample of outputs – checking whether cited documents or cases exist, say what the AI claims, and support the proposition for which they're cited. This can be a great learning exercise for interns and associates!
- ***For accuracy:*** Run a comparative study between your AI tool and a more traditional review workflow. Evaluate completeness, where the AI and previous reviewers disagreed, and what those results mean in terms of risk or insight.
- ***For productivity:*** Run a time-spent comparison on a defined task, with and without the tool, across a representative sample of real work. This kind of test is great during a pilot, before you're fully bought in to a new tool; the vendor can help you so it's not all on your team, you'll evaluate the tool's usefulness right away, and it will establish a benchmark to use over time if you do fully onboard it.
- ***For outcomes over time:*** Track which matters used AI assistance and at what stages, and log outcomes consistently. Longitudinal trending is a big evaluation to run, but it's also one of the most valuable. If you build this data-logging into your AI workflows from the start, Future You will be grateful! Here again, an AI assistant can be super helpful in conducting quick, regular analyses less expensively.

The Context and Calibrate stages will likely require the most discussion before you write down your final answers; embrace it. And make sure you involve all stakeholders; their input on things like risk appetite, testing capacity, budget, and more will be essential to a holistic evaluation strategy.

"Wherever on the spectrum of validation you fall – between rigorous statistical validation and maybe subjective user satisfaction – you need to know where you are and be on the same page with your partners to effectively evaluate your outcomes," says Mahendra Marshall.



Stage Four: Conduct

The Conduct stage encompasses the methods you use to test your claim in the context you've defined, at the rigor level you've calibrated.

This is where the actual testing happens. Fair warning: sometimes it doesn't go exactly as you'd expect, and the temptation to skip or shortcut can be strong.

There is no single right method, which might only make all the options in front of you feel more intimidating. For guidance on where to start (and how to keep things achievable), here are several practices that experienced legal teams return to consistently.

Test Against Known Answers

Angela O'Neal, a director at Nextra Solutions and an AI Visionary, offers some golden advice: "AI only works if you ask the right questions. A good way to test it is to ask something you already know the answer to."

Known-answer testing sounds simple, and it is! Which is exactly what makes it so effective as a starting point. If a system can't answer a question you already know the answer to, you have no reason to trust it on questions you don't.

This might look like:

- Running an e-discovery tool on data from a previous, finished case. Compare the AI's conclusions to those from your original workflow. Did it miss anything? Did it find anything new?
- Testing a drafting tool on itself. Give it a brief or motion draft it previously generated and which you've already looked over. Then, ask it to review for output accuracy and quality and check its answers against your own validation.

Run Parallel Processes

Before replacing an existing workflow with an AI-assisted one, run them side by side for a while. (This, too, is an excellent use of a new tool's pilot period.)

Compare the outputs as a team. Note where they agree, where they diverge, and the level of risk introduced by those divergences. Measure and document differences in time or money spent (across all stages, including research, generation, iteration, and verification), team sentiments around their user experience and perceived quality of outputs, and the like.

The parallel run will help you establish whether a potential tool is worth your while, or an updated version of your existing tool is still serving you as it should. If successful, this run will also provide your baseline to measure against moving forward.

Conduct Ongoing, Deliberate Sampling

For many use cases, you don't need to review every AI output – but checking on outputs overall is a wise idea. Sampling for quality control can be hugely beneficial to ensure a given tool's reliability and your team's effective use of it.

Building ongoing, randomized validation like this into your AI governance practices is a great way to build defensibility and optimization into your tech strategy. But remember: you've got to get the methodology right to get helpful results. Make sure your sample is representative of the range of inputs the system will encounter in real use, not just easy cases.

Keep Humans at Every Sign-off Point

Mahendra Marshall's principle applies here with full force: humans are accountable by default, and that accountability cannot be delegated to the system being evaluated.

Whatever methodology you use, the human judgment layer needs to be genuinely engaged. Real, knowledgeable people should be meaningfully reviewing, questioning, and verifying conclusions at the points where stakes are highest.



Stage Five: Continue

Evaluation done once is evaluation already expiring.

After you've put in all the effort to build and execute on an evaluation exercise, it's super easy to breathe a sigh of relief and check the task off your list. But AI tools evolve so rapidly that a system you validated six months ago may be operating on a different underlying model today. The evaluation done during onboarding or after a set project is a starting point, not a lifetime certification of reliability.

Aurora Bryant frames this as a strength rather than a burden: "Evaluations are an enduring asset. They're so critical – not just during an initial product development or onboarding lifecycle, but at any point where we're introducing change."

The Continue stage has two components: documentation and cadence.

- **Documentation** is about capturing what you tested, how you tested it, what you found, and what decisions you made as a result. This record has practical value (when a client or court asks you to show your work), professional value (demonstrating your diligence and commitment to ongoing learning), and operational value (as the baseline for your next evaluation).
- **Cadence** means establishing, in advance, when you'll re-test, and what changes in your tools, data, or use cases will trigger an earlier review.

Teams who build evaluation into their regular workflows as a standing practice develop institutional knowledge that will help them use AI systems safely, at scale, and over time. They're also best positioned to answer the question any client, partner, or court may eventually ask: ***How do you know it worked?***

Section 5: Evaluation Insights from the Field

The accumulated judgment of practitioners who have already run these evaluations is invaluable when you're orchestrating your own AI strategy.

Fortunately, the legal data intelligence community is a welcoming place to connect with and learn from others. Nurturing and engaging with a strong network within and beyond your organization will help you accelerate your learning curve and avoid common pitfalls.

To that end, make it a priority to attend webinars and **industry conferences**, engage in discussion on LinkedIn and other platforms, and read voraciously.

In the meantime, here are some insights and perspectives courtesy of our community.

©Relativity. All rights reserved.

How Practitioners Choose Tools

Before any evaluation methodology kicks in, you've got to know: how do practitioners decide which tools are **worth evaluating seriously** in the first place?

Pretty consistently, the answer starts with the vendor – and, more specifically, with how much transparency they're willing to offer about their own evaluation work.

Mahendra Marshall puts it plainly: "When it comes to the technology vendor, we do have an expectation of ongoing evaluations and as much transparency about those evaluations as possible. We need clearly defined goals and clearly defined measures of success that we're on the same page about – and maybe a level of protection against unreasonable expectations about the application, about the capability or suitability of systems."

Consider this part of your due diligence posture. Practitioners who have been through AI tool evaluations know that vendor capability claims and real-world performance can diverge significantly, and that misaligned expectations are a common reason that software adoptions fail.

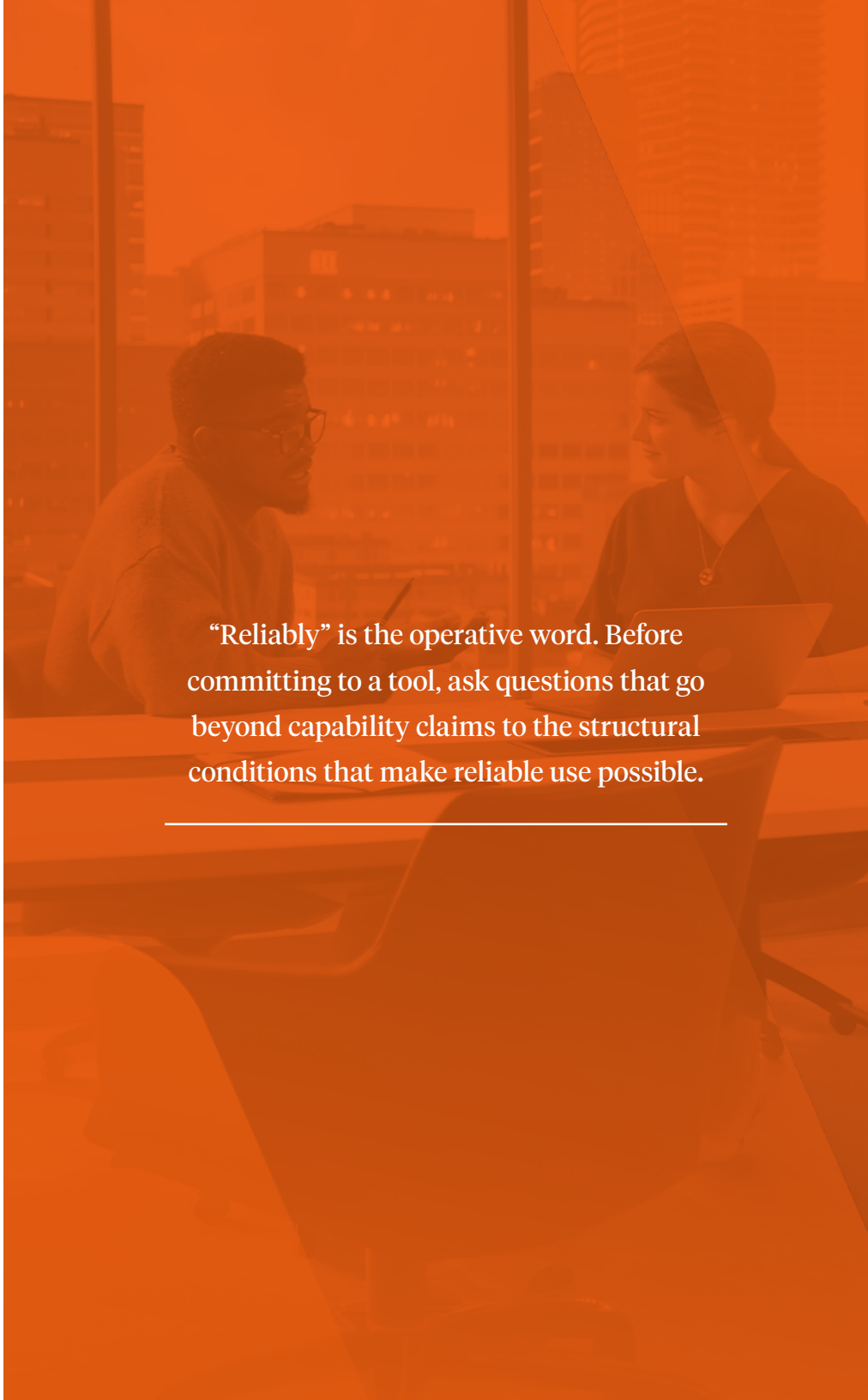
When technology evolves so rapidly, it's a best practice to look at your technology providers like partners rather than simply vendors.

"We have to be working together with a technology provider," Rishi Chhatwal explains, "because we have to understand what the background is and what the whole setup is. If we don't have that level of understanding and comfort, it's very hard for us to sign our names next to a new tool."

Aurora Bryant, on the vendor side of that conversation, describes what responsible transparency looks like from within a software provider: "Our role is building and delivering a product that is capable for the use cases that it was designed to do reliably. Not just once, not just twice, not just occasionally, but consistently at a level that meets the user's expectations."

"Reliably" is the operative word. Before committing to a tool, ask questions that go beyond capability claims to the structural conditions that make reliable use possible.

"Reliably" is the operative word. Before committing to a tool, ask questions that go beyond capability claims to the structural conditions that make reliable use possible.



©Relativity. All rights reserved.

Questions to Ask Any AI Vendor Before Adoption

- Do you meet our requirements for industry- and geography-specific compliance and regulatory standards and qualifications?
- What is your security posture? Who will have access to our data in your platform, and what is your incident response plan in the event of a breach?
- Can you contractually and technically switch off external model training on our data?
- What is the foundation model in your AI system? How did you train it for this use case?
- What does your audit trail look like? Can I access logs of users' activity, the system's inputs and outputs, prompts, and so on?
- What happens when the system doesn't know an answer, or produces a wrong answer? How will we know?
- What built-in validation and evaluation features does your system include?
- What does your team's evaluation and testing methodology look like, and can you share your findings?
- What does your release cycle look like? How do you communicate changes to the underlying model or system behavior?

As Ankur Malik, who shared many of these questions with us, puts it: "If a vendor cannot answer your question directly, that tells you what you need to know."

AI Visionary Gary Foster, managing director at Alvarez & Marsal, frames some starting discipline that can help you discern what's what.

"There are many ethical obligations legal technology professionals have when it comes to building, training, or using artificial intelligence in their work. The keys are privacy, transparency, and fairness. We work through these requirements closely with our partners who provide us with the underlying AI models we use," he explains.

Gary, like Ankur, suggests asking direct questions of any vendor to get to the heart of their claims.

"When it comes to using the model and the output provided, transparency is the key for us. We want clear reasoning behind the model's answers, and we make sure users are fully aware that the answers are model-based and inform them of the risks that may entail," Gary notes.

Of course, the issue of traceability holds weight legally as well as technically. If AI makes a claim you can't get the reasoning behind, it isn't a claim you can defend in front of a client, a judge, or your colleagues. Read vendor language the way you'd read a contract: not only for what it says, but for what it doesn't.

Understanding Risks Beyond Accuracy

Most AI evaluation conversations focus on accuracy, which you might use your 5C Framework to test via recall rates, error rates, hallucination frequency. These exercises are vitally important in many contexts.

But practitioners consistently identify a second category of risk that accuracy metrics don't necessarily capture: the human and organizational conditions that determine whether AI is used well.

Director of Legal Informatics at Gilbert + Tobin and AI Visionary Adrian Agius names it for us: "The risk isn't in experts doing the wrong thing; we have professional obligations to manage that concern. The risk is in well-intentioned people doing the wrong thing confidently. Our obligation is to design systems and education that make that harder, not easier."

Data-driven evaluation frameworks are essential, but behavioral and training best practices are what help prevent a competent professional from accepting a wrong answer because the output looked authoritative and well-formatted and they got it fast.

This is one reason that documenting and sharing evaluation results internally is so important. You dispel the notion that AI is either always right or always wrong, and aid in the AI fluency of your team by sharing lessons learned and the impact they have.

Another risk category is relational: in a law firm context especially, AI adoption decisions have to be made with the client in the proverbial room.

Rishi explains that, for Redgrave, "the client has to be involved in those conversations, and they have to know when what they're asking for may not be a good use case for that particular AI or that particular workflow."

Ray Mangum – a colleague of Rishi's – is a partner at Redgrave as well as an AI Visionary. He also identifies client alignment as a friction point that distinguishes law firm AI adoption from in-house adoption.

"The unique barrier to AI adoption in law firms is that it's not just up to you. The client has to be comfortable and aligned with your use of AI and their data. Clients are understandably concerned, and there are legitimate risks that must be managed," he explains. Explaining your evaluation strategies and showing your results will help address those concerns and optimize your use of AI for that client's unique matters.

Evaluation, in other words, is both a technical and a communication exercise, involving colleagues, clients, partners, the courts, and regulators who will conduct or eventually scrutinize the work it's measuring.

> "There are many ethical obligations legal technology professionals have when it comes to building, training, or using artificial intelligence in their work. The keys are privacy, transparency, and fairness.

©Relativity. All rights reserved.

Section 6: What the Industry Hasn't Settled

We all agree that evaluation is good, necessary, and worth doing well. But we've also noted that how to "do it well" is an uncomfortably open question for legal AI users.

There are other open questions on this topic, too, that are worth your attention. Here are a few that are in active discussion.

A good look at these debates will help you sharpen your evaluation mindset and arrive prepared for tough conversations around AI.

©Relativity. All rights reserved.

AI Evaluating AI

This is one of the more contentious live debates in the field. Both sides have merit, but discussions carry on.

The case for it: manually checking every output an AI system produces doesn't scale, and using a second AI model to flag likely errors, score outputs, or triage which results need expert review is a legitimate way to extend a human evaluation team's quality-control reach. It's already happening across the industry, including in legal-specific contexts where AI-based scoring is used to evaluate other AI-generated outputs for relevance, accuracy, and quality.

The case against it: an AI judging another AI's output may not deliver the confidence people look for when they're testing … AI. Shouldn't human experts be the ones to decide whether AI is reliable, rather than AI itself?

Multiple methodologies help data science teams grapple with this balance while leveraging AI's efficiency gains during their evaluation workflows.

For example, "there's a really common LLM-as-judge evaluation type that's called reference-free," Aron Ahmadia explains. "You give the LLM an output from a question-answering system and you say, 'Hey, is this any good?' The LLM only has the context you're giving it, and you can't ask the most important question, which is: 'Is it right?'"

The idea is to minimize bias or suggestion and see what the AI uncovers.

Aron's team at Relativity uses a bit of a different strategy: a "reference-based" approach.

"It's basically a graded rubric where we provide the question and we also provide the answer. At Relativity, we hired an entire team of lawyers to ensure the specs and standards meet these expectations."

This methodology gives the Relativity team a better frame of reference for the AI's grading and how it came to its conclusions.

For their team, using an LLM as a judge during evaluation and product development lifecycles helps them keep pace with the speed of innovation.

"Say I have a model behaving unreliably and I give it to a channel partner testing the tool to ask, 'How's this doing for you? What's it doing well and what's it not?' They might give me a ton of feedback, but it might be a week or two before I get that answer. If I have an LLM-as-judge system in place, I get its answer within seconds," Aron says. "I can incorporate that into a fully automated system that provides feedback continuously. So in the same way that a tool like aiR for Review can save you thousands of hours of lawyer time in evaluating documents, an internal LLM's judge system can save users from thousands of hours of troubleshooting, because we can run it through a thousand times and get high-quality feedback."

To sum up: AI-assisted evaluation is a reasonable tool for triage and scale, but it is not a substitute for the grounded, reference-based, or human-reviewed verification that legal use cases demand.

Tolerance of Output Variability

Legal systems want certainty. Rishi Chhatwal put it plainly: "Everybody wants 100 percent – but I just don't know how reliable that would actually be."

There's indisputable tension between this profession's appetite for guarantees and the probabilistic nature of how these models actually work. Unfortunately, that tension is not likely to go away soon, and the industry is sort of stuck grappling with it.

Again, though, context helps. e-Discovery has decades of precedent for validating non-deterministic tools: technology-assisted review has long relied on recall, precision, and elusion metrics to establish that a process is defensible without requiring it to be statistically perfect.

©Relativity. All rights reserved.



Generative AI for doc review, for its part, maps nicely to these metrics, and legal data review teams are already using them to build defensible ESI protocols and workflows around these systems.

In other use cases, though, there often isn't a definition of what "statistically defensible" looks like. You can evaluate whether outputs feature correct answers, accurate citations, sound arguments, and so on. Of course this is necessary validation, but there remains some nuance. For instance, we know that, if you give an AI drafting or research tool the same prompt twice, the outputs may differ quite a bit even if they're both correct. It's tough to point to a stat or equation that can prove which is objectively better or whether those differences are too great.

The Claim and Context stages of the 5C Framework help you level-set. "How much variability is okay" doesn't have one answer; in fact, it has as many answers as you have use cases. But you can work with that.

The Necessity of Shared Industry Evaluation Standards

This unresolved debate is playing out in real time at the institutions we all turn to for professional guidance.

The **Sedona Conference's Working Group 13** – a body specifically formed to coordinate AI-related work across the legal industry and recommend principles and best practices – is currently in the middle of a structured effort built around three workstreams: engineering, evaluation, and drafting. Its evaluation workstream's white paper isn't expected for member comment until later in 2026. While one of the industry's most credible bodies to propose legal AI evaluation standards is still actively drafting them, many eagerly await their perspectives.

EDRM is moving on a parallel track, **expanding** its Information Governance Reference Model to address AI-era governance and naming the stakeholder roles that need a shared vocabulary. It's a governance framework, rather than an evaluation protocol, but still a helpful reference for teams looking to adopt AI thoughtfully.

ISO has published broader AI standards, including a **2025 framework for AI system impact assessments** and a **taxonomy for AI transparency**, both of which present useful scaffolding. Neither is legal-domain-specific, and neither tells a litigation team how rigorously to test a privilege-review model (for example) before relying on it, but their guidance is insightful from a broader mindset.

In the UK, the **Civil Justice Council's working group** is consulting specifically on whether procedural rules are needed to govern AI use in court documents.

Sitting in this development period, watching the industry adapt in real time to some truly transformative technology, is exciting. But waiting for guidance can be troublesome, too: because innovation happens a lot faster than rule updates, everyone's competition is finding new ways to use AI to their advantage, and the dichotomy of "give it a try before we fall behind" and "wait for safe instructions" is not easy to sit with.

On one hand, standardization would offer consistency and a shared baseline of trust.

On the other, the importance of context-specific validation argues that a model evaluated for contract review and a model evaluated for privilege determinations may simply need different tests anyway.

Both of those statements are true, and the industry hasn't decided how to reconcile them. At the current pace of change, though, waiting for them to figure it out may be less beneficial than building a thoughtful evaluation discipline that can help you find the right balance for your current needs.

©Relativity. All rights reserved.

Section 7: Planning for the Future

By now, you've internalized the reality that evaluation is more about infrastructure than simple exercise.

A one-time vendor bake-off tells you whether a tool was good enough on the day you tested it, but very little about whether it's still good six months later, after a model upgrade, following a shift in case law, or with a new patch from the developer.

Mahendra Marshall emphasizes that "the pace of change is obviously incredibly rapid, and I do think it stresses the importance of ongoing evaluation."

Rishi Chhatwal extends the point in two directions: "The pace of change is very fast in both the technology outside the legal field – some of which we rely on for foundation models – and also in the acceptance that is happening inside the legal field. Today's answers are not tomorrow's answers."

Your evaluation strategy is an asset you build. For better or for worse, it compounds in value the more consistently you maintain it and depreciates the moment you stop.

©Relativity. All rights reserved.

Evaluation as a Defensibility Record

There's a forward-looking implication here that you should try not to miss: the evaluation work you do today may itself become evidence later. As AI becomes embedded in legal workflows, its prompts, chat logs, and model outputs are becoming matter artifacts in their own right – material that may be discoverable, reviewable, or scrutinized after the fact.

Under such circumstances, the competitive question shifts a bit. The edge goes to whomever can produce a clean, contemporaneous record of how a tool was evaluated, what it was found capable of, and what guardrails were put around its use.

Rather than a compliance afterthought or dull governance policy, AI evaluation proficiency may become a genuine differentiator for legal teams who want to stay ahead of the innovation curve without putting their defensibility or professional obligations at risk.

Evaluation Maturity Reflects AI Maturity

Organizations tend to move through AI evaluation in a predictable arc:

- Ad hoc spot-checks first, where someone notices a wrong answer and flags it informally.
- Then: defined validation workflows, where evaluation criteria, cadence, and ownership get written down.
- And eventually: instrumented, continuously monitored systems, where evaluation runs in the background as a function of how the tool operates rather than as its own project.

Given the adoption and ROI research we discussed earlier, most of our industry is somewhere in the first or second stage. All over the community, legal teams like yours are working their way toward greater AI maturity.

As your AI fluency increases, so will the strength and soundness of your evaluation strategy – and that's the way it should be. These functions work best when they're working together.



Your evaluation strategy is an asset you build. For better or for worse, it compounds in value the more consistently you maintain it and depreciates the moment you stop.

©Relativity. All rights reserved.

Appendix: Related Resources to Keep Exploring

Looking for more hands-on information about how to adopt, advocate for, and evaluate AI? These resources may help.

Defensible Validation of Relativity aiR for Review: Metrics, Methods, and Considerations

Download this white paper to learn how you can easily validate performance, avoid costly missteps, and bring defensible generative AI into your review workflow from here on out.

The White Paper | Blog Post

Putting Gen AI to the Test: A Document Review Accuracy Study

This study used a low-richness data set requiring a nuanced legal standard to evaluate a generative AI review workflow against a managed review using active learning. The results show that generative AI can meaningfully improve both efficiency and review completeness – while supporting human decision-making in complex matters.

The Study | Blog Post | Interactive Game

Drawing Conclusions: Negotiating AI in Your ESI Protocols

How do you balance innovation with respect for professional caution when the rules are still being written? This delightfully illustrated e-book imagines a mock meet-and-confer debate on the use of generative AI for e-discovery.

The Storybook | Blog Post

The Relativity Blog

Want an easy way to stay on top of legal tech trends? Subscribe to ***The Relativity Blog*** to receive guidance on optimizing your team's legal data intelligence and AI strategies, commentary and perspectives from industry pros, and more.

Subscribe

Legal AI Innovation Newsletter

Follow along with legal AI news, educational content, and product announcements from the Relativity community and beyond.

Subscribe

Relativity aiR

From technology-assisted review to agentic workflows, Relativity has been leading AI innovation for over a decade. As legal data grows and change accelerates, we're moving with purpose – partnering with our community to deliver useful, principled AI built to tackle the thorniest legal data challenges.

Learn More

Relativity
231 South LaSalle Street, 20th Floor • Chicago, IL 60604
+1 (312) 263-1177 • relativity.com

©Relativity. All rights reserved.